DOCUMENT RESUME

ED 349 318 TM 018 917

AUTHOR Franklin, Jennifer; And Others
TITLE Grade Inflation and Student Ratings: A Closer Look.
PUB DATE Apr 91
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Class Size; *College Faculty; *Courses; Evaluation Methods; *Grade Inflation; Higher Education; Longitudinal Studies; Predictor Variables; Scoring; *Student Evaluation of Teacher Performance; *Teacher Effectiveness; *Undergraduate Students
IDENTIFIERS *Frequency of Measurement

ABSTRACT

Although many sources have discounted the likelihood of grade inflation resulting from instructors trying to "buy" better student ratings of instruction, many faculty members still believe that there is widespread manipulation of grades. This study involving 28,629 undergraduates over a 6-year period found no relationship between frequency of evaluation and 2 indices of grade inflation: average end of term grades and student ratings of course workload. Participation levels evaluated ranged from fewer than 500 courses per quarter to over 1,300. Student ratings of course difficulty tended to be higher for more frequently evaluated instructors. No differences in grading patterns over the 6-year study period for instructors were observed for changes in evaluation contingencies such as whether ratings were voluntary and confidential or mandatory and published. Class size proved to be an important predictor both of ratings of instructor effectiveness and end of term grades. Six tables present study data, and a 20-item list of references is included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
MICHAEL THEALL
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED338038

Grade Inflation and Student Ratings: a Closer Look

Jennifer Franklin
Northeastern University

Michael Theall
University of Alabama, Birmingham

Larry Ludlow
Boston College

Although many sources have discounted the likelihood of grade inflation resulting from instructors trying to "buy" better student ratings of instruction, many faculty still believe that there is widespread manipulation of grades. This study found no relationship between frequency of evaluation and two indices of grade inflation: average end of term grades and student ratings of course workload. Student ratings of course difficulty tended to be higher for more frequently evaluated instructors. No differences in grading patterns over the six year study period for instructors were observed for changes in evaluation contingencies such as whether ratings were voluntary and confidential or mandatory and published. Class size proved to be an important predictor of both ratings of instructor effectiveness and end of term grades.

Office of Instructional Research and Evaluation, 305 CU
Northeastern University
360 Huntington Ave.
Boston, MA 02115
(617) 437 4896

Paper presented at the annual meeting of the American Education Research Association, Chicago, April 1991.

Please do not cite or reproduce without permission of the authors.

TM018127

BEST COPY AVAILABLE

Grade inflation and student ratings: a closer look

Objective: This study investigates the relationship between final course grades and student ratings (workload, difficulty, and overall instructor effectiveness); at which interval in a six year period student ratings were obtained; and how often ratings were obtained. The purpose of the study is to detect shifts in grading practices and/or in students' perceptions of workload and difficulty which could indicate the presence of grade inflation as a result of an instructor's anticipation of being rated by students (a.k.a. teacher course evaluation or T.C.E.)

Theoretical Perspective: It is commonly accepted that the findings of multisection validity studies (Cohen, P., 1981) have resulted in widespread acknowledgment of some moderate degree of positive association between real student achievement and students' ratings of their instructors and courses. This association has become a cornerstone of the case for the validity of TCEs: good ratings reflect students' satisfaction with their own achievement which is a least partly the result of the teaching they have received. Nonetheless, research and practice literature sources remain inconclusive on the subject of grade inflation, sometimes referred to as grading leniency. That giving grades that are higher than student achievement warrants or lowering course standards can "buy" higher student ratings is alleged in a variety of published opinion pieces on ratings (e.g., Martinson & Ryan, 1984; Machina, 1987; Myers, 1979). Ratings are have been blamed as a major culprit in what some claim to be a general trend toward grade inflation and lower standards in higher education. (e.g. Goldman, 1985, Cohen, H., 1984)

A number of studies have attempted to detect grade inflation or instructor leniency in the context of student ratings using students reports of expected grades, grade point average, and (infrequently) actual grade distributions. Grade inflation in the context of student ratings is operationally defined variously as the biasing effect of students expectations of unearned high grades, increased leniency with respect to workload or difficulty level, grading leniency, or in economic models, *quid pro quo*. (for a particularly interesting example of the latter, Nelson & Lynch, 1984).

Grades are a complex product of many student, instructor, and situational variables. Moreover variation in the standards, methods, and philosophies of grading makes "grade point average" a questionable measure for many purposes. Despite many methodological problems in comparing grade point averages and other indices of "grade inflation", with other measures of achievement, studies of postsecondary grade distributions have detected occasional shifts in grading patterns and attribute them to a variety of circumstances. (e.g., instructors responding to institutional pressure or economic concerns based on declining enrollments, personal crises of students, motivational factors based on recognition of student effort). (Milton et al, 1986)

In studies done to determine the validity and reliability of student ratings, grades or expected grades were frequently used as a measure of student achievement. A few examined the effects of discrepancy between expected grades and grade point average (G.P.A.) as a source of bias in ratings. Feldman (1976) looked at over 200 studies and was unable to conclude either that grades bias ratings or that grades could be ruled out as a source of bias. Aleamoni and Hexner (1978) cited 22 studies finding small negative or no associations between ratings and measures of achievement and 28 finding modest to moderate positive associations. It must be noted that in the majority of studies considered, grades themselves are based on various and usually unvalidated achievement measures, frequently confounded by the fact that they have been constructed by those being rated and giving the grades (Costin, 1978).

Practical response to the ambiguous state of the literature varies. Several knowledgeable sources have suggested that grade distributions should be examined along with ratings for personnel decision-making or else have voiced concern that grade-induced bias in ratings is possible (e.g. McKeachie, 1979; Centra, 1979; Doyle, 1975). Meanwhile, others have counted the notion among the popular "myths" associated with ratings or else discounted its importance, even when conceding the possibility of its occurrence (e.g. Costin et al, 1971; Aleamoni, 1987; Seldin, 1987). Marsh, in his 1987 review, considers the effect of grading leniency on ratings inconsequential when observed but, otherwise, inconclusive and notes elsewhere that there may be a conceptual problem associated with calling the effect of certain variables "bias" when substantive relationships between these variables and achievement might be occurring.

However, it is difficult to draw conclusions about grade inflation based on observed associations between ratings and expected or actual grades when grades and ratings are each measures of student satisfaction and teacher performance. No studies have reported investigation of the long range effects of increased ratings use at

the institutional level. Ultimately, changes in grading patterns or lowering of the effort required to earn grades must be demonstrated to make the case for a systematic relationship between grade inflation and ratings. This shift would logically occur when faculty expect to be rated if the goal is to effect improvement in ratings, *quid pro quo*.

Whether grade inflation occurs as a consequence of the expectation or experience of being rated by students does not necessarily depend on whether raters are actually influenced by grades or leniency in course standards. Whether or not giving unearned good grades improves ratings, many faculty believe it does and that "someone" is doing it. (Franklin & Theall, 1989) When ratings are used for merit review and affect professional reputations, it is reasonable to consider whether some faculty may have motive to inflate grades or lower course standards in the hope of improving ratings. Without systematic and periodic administrative review of grade distributions, they may also have the opportunity. Changes in grading standards could occur and not be readily detected.

Grade inflation caused by collecting student ratings would be a serious threat to fair and valid practice of postsecondary personnel decision-making. It would invalidate the consumer-related purpose of providing ratings results to students for course selection and put a chill on the use of ratings for teaching improvement as well. Rumors of its existence also provide a powerful disincentive for participation in TCEs or taking ratings seriously. Thus, even though most sources at least agree that grades (expected or given) do not play a major role in ratings outcomes, it is still important to demonstrate that systematic changes in grading or course standards are not occurring. It might be easier to make the case that ratings do not cause grade inflation by demonstrating that no changes in grading have occurred than to persuade skeptical faculty that a positive association between ratings and grades is actually evidence for of the validity of ratings.

This study focuses on what effect being evaluated by students has on instructors' grade-giving behavior rather than what effect being graded has on the students' evaluations of instructors or courses. The study takes a longitudinal approach to the problem of detecting grade inflation by looking over a six year period for systematic changes in the distributions and means of grades given by instructors with respect to their expectation or the actual event of being rated. Since this study looks at patterns of grade distributions as a dependent variable, the fact that the instructors who were rated also assigned the grades is essential to the study rather than a source of confounding as in previous studies.

Method: We made no apriori assumptions concerning the existence and prevalence of grade inflation. We assumed only that it would be more likely to happen when instructors were anticipating or actually being rated in any course-section. We also hypothesized relationships among a number of contextual variables (departmental affiliation, competing departmental evaluations, student ratings of overall instructor, overall amount learned, overall course, size of class, and average grades for rated and non-rated courses).

We tested the following hypotheses:

A. H_o There is no relationship between how often courses are rated and the mean of grades given in those courses.
   H_1 The more sections of a course that are rated, the higher the mean of grades given in rated sections compared with those given in unrated sections of the same course.

B. H_o There is no relationship between how often instructors are rated and the mean of grades given by those instructors.
   H_1 The more sections rated compared with the number of sections taught for a given instructor, the higher the mean of grades given by that instructor.

C. H_o There is no relationship between how often courses are rated and mean ratings of those courses.
   H_1 The more often courses are rated, the higher the mean ratings.

D. H_o There is no relationship between how often individual instructors are rated and mean ratings for those instructors
   H_1 The more often an individual instructor is rated the higher the ratings

E. H_o There is no relationship between how often courses or instructors are rated and mean ratings of workload and/or difficulty.
   H_1 The more often courses or instructors are rated, the more mean ratings will indicate lighter workload and/or less difficulty.

F. H_o There is no relationship between the size of course-sections and the mean of grades given in those sections
H_1 There will be a decrease in the mean of grades given as the size of course-sections increases.

G. H_o There is no relationship between the size of course-sections and the mean ratings of those sections
H_1 There will be a decrease in the means of ratings as the size of course-sections increases.

H. H_o There are no differences in the means of grades given in courses where evaluation is mandatory and/or results are made public and those couses where evaluation is voluntary and/or results are confidential.
H_1 Means of grades from courses where evaluations are mandatory and or whose results are public will be higher than means of grades from courses where evaluation is voluntary or results are confidential.

Subjects: The study population includes all undergraduate day division course sections taught at a large private, urban university between fall 1984 through spring 1990 for which a valid grade point average for the course had been recorded (n=28,629). Courses span a broad range of content and discipline areas (10 colleges, 60 departments). Of these courses, 13,090 course sections participated in a centralized evaluation service system which collected TCEs each Fall, Winter, and Spring quarter. Departmental participation ranged from 20 to over 90 percent. Because campus-wide ratings were mandated by the Faculty Senate, but were established incrementally using field tests to validate instruments and procedures, participation levels over the six year period range from fewer than 500 courses per quarter (summers excluded) to over 1300. (Table 1).

Courses were evaluated using a validated 40 item diagnostic questionnaire, its subset, a 10 item summary questionnaire, or a 40 item precursor to the diagnostic form. The 3 items reported in this study ("instructor's teaching effectiveness", "workload", and "difficulty") were included in all three forms. Results of 10 item questionnaires were always published annually in a catalog of results. Results of diagnostic items were always confidential, but the publication of the 3 items varied during the six year period. Results for the 3 summary items from one academic unit (approximately 1/5 of our sample) were automatically reported to its department chairs. All ratings were obtained anonymously from students, using a standard administration procedure (volunteer student monitors administering, collecting, and returning forms to independent scoring service; instructor out of the room, etc.).

Each course-section in the sample can be classified in terms of both history and purpose for obtaining ratings. In the "pilot" phase, partipation was voluntary and only Business School "overall items" were published. In the "33 1/3" phase only one third of undergraduate courses were selected for mandatory evaluation each quarter. In the "100%" phase, all undergraduate courses were selected for mandatory evaluation. In each of the three phases, faculty were notified of the TCE selections during the fourth week of the quarter and given the opportunity to request the diagnostic questionnaire (with summary items extracted for publication during the second two phases).

Ratings of all types were obtained in an additional 652 courses but were not included in analyses since these results were invalidated for errors in administration such as failure to use a student monitor. The total of non-participating course-sections was 10,333. Reasons for non-participation included simple non-compliance, official exemption granted for one or more of a variety of technical reasons, or that the instructor teaching a particular course-section had not been included in that term's TCE target group.

Course-sections of courses which had one or more valid TCEs conducted during the study period were chosen. Course-sections in which TCEs were attempted, but invalidated due to errors of administration were excluded from the study. Course-sections with multiple instructors, enrollments under 5, half-term courses, and other circumstances which might invalidate TCE results had been formally exempted

Table 1. Study Sample by Academic Year

Academic Year	LISTED SECS	TOTAL TCES
Pilot phase:		
FA 1984 - SP 1985	4680	837
FA 1985 - SP 1986	4594	1609
33 1/3% TCE sampling phase		
FA 1986 - SP 1987	4549	1440
FA 1987 - SP 1988	4497	1169
100 % TCE phase		
FA 1988 - SP 1989	4665	2952
FA 1989 - SP 1990	4685	2904

from the TCE process.

Three levels of analysis were used: course-section, courses aggregated by course number, and courses aggregated by instructor. The final selection criterion for courses, instructors, and course-sections required any selected course-section to be from a section of a course having at least one section with a valid TCE. Only course-sections meeting the criterion were aggregated for further study. (Average course grades for both courses and instructors meeting the selection criteria but with no history of participation were also examined.)

The general strategy was to study associations among the study variables and test for differences for variables hypothesized to cause changes in grades or the effort required to earn them. Grades include: the mean of grades given for each course-section in sample, the grand mean of all course section grades for each course, and the grand mean of all course section grades given by each instructor. Ratings include: the mean of each of the 3 ratings items obtained in each course-sections in the sample, the grand mean of all course-section ratings of the 3 items for each course, and the grand mean of all course-section ratings received by each instructor. The ratings items (from the TCEP Questionnaire) included: overall rating of the instructor 's teaching effectiveness, course workload (light to heavy), and course difficulty (easy to difficult). Ratings for two additional items, "overall course" and "amount learned", are reported in Tables 2, 3, and 4. Independent variables included: academic year; the percentage of course-sections evaluated; and groups reflecting the changing basis of evaluation contingencies over the six years, e.g. voluntary vs mandatory; published vs confidential. Because previous validation studies of the ratings instrument indicated class size to be a source of systematic variation in ratings, class size was included in this study. Based on range tests, 3 categories of class size were established: under 20, small; 20 - 49, medium; over 49, large). Class size based actual enrollment (interval) was also used as a control variable in correlation studies of the data.

Results: Initial analyses on these data found significant group differences in both ratings and grades varying according to academic year and class size. The direction and strength of results warranted further investigation. However, further analysis demonstrated that several groups did not meet tests for homogeneity of variance. For those analyses and groups, data were transformed using the square root method. A series of one way analyses of variance, range tests, and correlational studies were performed. The results of the correlational analyses appear in the tables on the facing page (Table 2). The analyses of variance are summarized in the Tables 5 and 6.

The percentage of times a given course offering or instructor was evaluated and the average of grades given were found to actually have a small, but significant inverse correlation. The association of frequent evaluation with lower grades at each level of analysis suggests the need to control for the relative frequency with which particular courses may be evaluated in subsequent studies of this data. At the course-section level, sections in which TCEs were not administered actually had higher average grades ($\bar{x}$=2.79) than those participating ($\bar{x}$=2.73) Similarly, when grades were aggregated by course, a small, but statistically significant difference in grades based on percent of sections evaluated appeared. It is probable that this means that some courses are more likely to be evaluated than others. The most important finding is that there was no difference in the grades given by instructors based on how many times they were evaluated over the six year period.

Associations found between actual grades and overall ratings of instructor, course, and overall learning were consistent with those found between expected grades and overall ratings both in this sample and in the literature. At the course and course section levels, significant differences between course-sections were found on the basis of workload, difficulty, and instructor ratings.

Class size emerged as the single most powerful predictor of grades and ratings for single-course sections, courses, and instructors. The relatively strong inverse correlations between class size and grades in each level of analysis may be the result of differences in grading standards, methods, or philosophy for small sections versus large sections; or a selection bias placing more experienced/higher achieving students (and, hence, more satisfied) in small, elective, or upper level courses. Similarly, the pattern of inverse associations found between class size and overall instructor ratings at each level of analysis suggest that the student's lack of satisfaction with the instructor is matched by a lack of achievement in larger sections compared with smaller ones.

The purposes for which ratings were collected were associated with differences in grades and ratings. Notably, course-sections in which the confidential diagnostic questionnaire was used (whether or not the summary items were published) were likely to be associated with the highest grades, receive the highest instructor ratings., and receive lower ratings of difficulty, but significant difference in workload ratings was found for these classes. It is tempting to speculate on whether instructors electing to use the diagnostic evaluation may be self-selected for their interest in improving their teaching skills and consequently provide more effective instruction. We have noted some "recidivism" in use of the diagnostic form among faculty in this sample.

TABLE 2. Product Moment Correlations Between Course -Section Level Variables (n=10,329)

		1	2	3	4	5	6	7	8
1.	Grades	1.00	-.27	.25	.24	.33	.24	.38	
2.	Class Size		1.00	-.10	-.12	-.12			.06
3.	Instructor			1.00	.84	.90	.05	.10	
4.	Learning				1.00	.89	-.08		
5.	Course					1.00	.08	.16	
6.	Workload *						1.00	.73	-.09
7.	Difficulty **							1.00	
8.	Pct Eval'd								1.00

TABLE 3. Product Moment Correlations Between Course-Section Level Variables Aggregated by Course (n=1599)

		1	2	3	4	5	6	7	8
1.	Grades	1.00	-.34	.26	.22	.33	.28	.34	
2.	Class Size		1.00	-.15	-.17	-.17		.05	
3.	Instructor			1.00	.83	.89		.07	-.09
4.	Learning				1.00	.89	-.12	-.05	.07
5.	Course					1.00		.11	.11
6.	Workload *						1.00	.77	.09
7.	Difficulty **							1.00	
8.	Pct Eval'd								1.00

TABLE 4. Product Moment Correlations Between Course-Section Level Variables Aggregated by Instructor Name: (n=2041)

		1	2	3	4	5	6	7	8
1.	Grades	1.00	-.28	.24	.25	.34	.29	.43	
2.	Class Size		1.00	-.09	-.13	-.13			
3.	Instructor			1.00	.87	.92		.11	
4.	Learning				1.00	.92	.09	.18	.07
5.	Course					1.00	.09	.18	
6.	Workload*						1.00	.76	
7.	Difficulty**							1.00	
8.	Pct Eval'd								1.00

Correlations reported in above tables are those greater than $r = .05$, $p < .001$

* workload item was scaled "one of the lightest" (high) to "one of the heaviest" (low)
** difficulty item was scaled "extremely easy" (high) to "extremely difficult" (low)

Table 5. Analysis of Variance Summaries for Grades

	Groups	*df*	*MS*	*F*	*Comment*
Course-section grade mean by TCE participation (yes or no)	Between	1	1.6988	6.8257 **	rated courses had lower grades
	Within	13078	.2489		
Course-section grade mean by academic year (FA 84 - SP 90)	Between	5	10.8486	37.0756 ***	grades increased over time
	Within	23407	.2926		
Course-section grade mean by TCE group (pilot, 33%, 100%)	Between	2	.4759	1.9708	
	Within	1034	.2415		
Course-section grade mean by workload	Between	341	.0904	4.3571 ***	higher grades, lighter workload
	Within	10659	.0207		
Course-section grade mean by difficulty	Between	279	.1638	8.4491 ***	higher grades, less difficult
	Within	10720	.0193		
Course-section grade mean *by instructor rating*	Between	332	.0822	3.9048 ***	higher grades, higher ratings
	Within	10668	.0211		
Course-section grade means by class size (sm, med, lg)	Between	2	213.6172	836.2910 ***	larger classes, lower ratings
	Within	18196	.2554		
Aggregated Course grade means by % TCE sections	Between	247	.2557	1.4687 ***	more evaluations, lower grades
	Within	1350	.1741		
Aggregated course grade means by instructor rating	Between	1048	.1967	1,1726 *	higher grades, higher ratings
	Within	549	.1677		
Aggregated course grade means by workload rating	Between	1014	.1956	1.1417 *	higher grades, lighter workload
	Within	583	.1713		
Aggregated course grade means by difficulty rating	*Between*	975	.2038	1.2734 ***	higher grades less diffficult
	Within	622	.1600		
Aggregated course grade means by class size (sm, med, lg)	Between	2	23.9873	152.8888 **	larger classes lower grades
	Within	1597	.0569		
Aggregated Instructor grade means by % TCE sections	Between	248	.1811	1.0267	
	Within	1749	.1764		
Aggregated Instructor grade means by instructor rating	Between	1310	.1775	1.003	
	Within	687	.1760		
Aggregated Instructor grade means by workload rating	Between	1226	.1805	1.0538	
	Within	771	.1713		
Aggregated Instructor grade means by difficulty rating	Between	1210	.1876	1.1675 **	higher grades, less difficult
	Within	787	.1606		
Aggregated Instructor grade means by class size (sm, med, lg)	Between	2	19.0635	120.6394 ***	larger classes, lower grades
	Within	1994	.1580		

* $p < .05$
** $p < .01$
*** $p < .001$

Table 6. Analysis of Variance Summaries for Ratings

	Groups	*df*	*MS*	*F*	*Comments*
Course-section instructor ratings	Between	5	.8071	1.9941	
by academic year (FA 84 - SP 90)	Within	10995	.4047		
Course-section instructor ratings	Between	2	.2119	.5277	
by TCE group (pilot, 33%, 100%)	Within	10304	.4014		
Course-section workload rating	Between	5	.5256	1.8916	
by academic year (FA 84 - SP 90)	Within	10995	.2779		
Course-section workload rating	Between	2	.0924	.3347	
by TCE group (pilot, 33%, 100%)	Within	10304	.2762		
Course-section difficulty ratings	Between	5	.5961	3.3198 **	'difficulty' ratings
by academic year (FA 84 - SP 90)	Within	10994	.1796		increase over time
Course-section difficulty ratings	Between	2	1.2053	6.7249 **	100% group
by TCE group (pilot, 33%, 100%)	Within	10303	.1792		highest difficulty
Course-section instructor ratings	Between	2	34.4641	86.4304 ***	highest ratings
by class size (sm, med, lg)	Within	10989	.3988		small classes
Aggregated Course Instructor ratings	Between	247	.1784	.6454	
by % TCE sections	Within	1350	.2764		
Aggregated course workload rating	Between	247	.1963	.7467	
by % TCE sections	Within	1350	.2628		
Aggregated course difficulty rating	*Between*	247	.1176	.6511	
by % TCE sections	Within	1350	.1806		
Aggregated course instructor ratings	Between	920	.0187	1.4184 ***	highest ratings
by class size (sm, med, lg)	Within	677	.0132		small classes
Aggregated Instructor ratings	Between	248	.1811	1.0267	
by % TCE sections	Within	1749	.1764		
Aggregated Instructor workload rating	Between	248	.2294	1.0836	
by % TCE sections	Within	1749	.2117		
Aggregated Instructor difficulty rating	Between	248	.1270	.8712	
by % TCE sections	Within	1749	.1458		
Aggregated instructor ratings	Between	2	19.0635	120.6394 ***	small classes
by class size (sm, med, lg)	Within	1995	.1580		highest ratings

* $p < .05$
** $p < .01$
*** $p < .001$

There is no evidence to support the existence of grade inflation based on a decrease in the amount of work or effort needed to earn grades. If anything, although workoad and difficulty are highly correlated (.73), course difficulty appears to have increased while the workload remained the same during the same period that ratings participation increased. As reported in most studies, workload appears unrelated to instructor rating at any of our three levels of analysis. However, a very small positive assoociation did appear between instructor ratings and students reports of the difficulty of the course work and materials.

A curvilinear relationship between class size and both workload and difficulty appeared in some groups within the sample, with workload heaviest and difficulty highest in small and large classes. Aggregated by course, students' self-reports of amount learned increased with higher workloads (-.12). Aggregated by instructor, students' ratings of instructor, course, and amount learned decreased with increased levels of difficulty.

The slight overall trend toward increase in average grades in the study population over the study period was of no practical importance, although it was statistically significant due to the size of the sample. Range tests (Scheffe at $p< .05$) did establish that the sample grade average ($\bar{x}$=2.78) was slightly more than .1 higher for the second three year period than the first ($\bar{x}$=2.68) but to connect this trend to the rise of ratings use would require the supposition that the effect of ratings so permeated the system that everyone simultaneously changed their grading practices or course standards.

Although this study covers six years of ratings and grades in a campus-wide TCE system at one of the largest private universities in the country, one particular limitation to this study is that it does not account for several local departmental evaluation systems which operated during the study period. At minimum, some departmental ratings policies reduced participation in the campus-wide service reported here. However, preliminary study suggests that ratings obtained in these departments appear to follow the same patterns as the larger sample. Other departmental and disciplinary differences not reported here appeared for this data. For example, the results found for the academic unit that had a mandatory evaluation policy over the six year span of the study (with nearly 90% participation, accounting for 20% of the courses reported here) were substantially similar to those obtained under all other departmental evaluation policy contingencies.

Conclusions: Had we found evidence for grade inflation associated with student ratings , we would still speculate that giving unearned good grades would not work very effectively for buying good ratings since ratings of instructors did not increase over the study period despite the fact that the number of TCEs conducted doubled in the last third of the study period. Yet, we can speculate about the basis for the intuitive appeal of the belief that some faculty inflate grades or lower standards to improve their ratings without improving the quality of the instruction they offer. If faculty observation and experience mirror the positive associations found in this study between higher grades and instructor ratings, lighter workloads, and lower levels of difficulty, respectively, it seems likely that many might construe such associations as causal relationships.

However, the various constituencies who collect, report, and use ratings data for personnel decision-making can perhaps rest a little easier since no evidence of systematic or pervasive changes in the actual grading behaviors of teachers was found based on whether they were rated or not or how often. Similarly, no evidence for the systematic lowering of course standards (workload or difficulty) was detected. The grading behaviors of faculty did not vary according to whether their ratings were published and hence readily available to administrators.

Although students' perceptions of workload have held constant over the study period, they reported more difficulty doing that work in recent years. Although instructor ratings appeared slightly higher for courses rated lower in difficulty, this is likely to be at least partly an artifact of content area differences in ratings. It may be helpful to remind faculty and administrative ratings users that workload and difficulty are not inherently good or bad. For example, it is possible for a poor instructor to give a lot of irrelevant work or offer instruction that is difficult because it is poor. Conversely, an effective instructor facilitates student achievement and, in a sense, makes learning easier. In any case, workload and difficulty appear considerably less related to ratings or grades than both ratings and grades are related to class size.

Although this study excluded graduate level courses, it did not control for the content area or level of undergraduate courses. Content area or discipline is known to be a source of variation in ratings and grades. It is also the case that the incentive to influence ratings would logically be affected by departmental personnel policy and instructor variables such as rank. Since differences in grades, ratings, and difficulty were observed based on class size, and since class size and the rank of the teacher assigned presumably vary according to content area and class level, more complex factorial approaches testing group differences and probing for interactions are needed. Despite the relatively large sample of course sections in this study, even more data is required to allow for sampling small departments.

Meanwhile, it might be more productive for ratings users to be concerned about known sources of bias such as improper or non-standard administration procedures. Despite the very popular and persistent myth that the "other guy" is "buying" unearned good ratings and that ratings are responsible for a decline in all things good in post-secondary education, not finding rampant grade inflation associated with ratings may help practioners assist their clients improve their interpretation and application of ratings to personnel decision-making, since many of the real-life consequences of ratings take shape in that domain. Teaching improvement consultants may also find these results useful to allay some of their clients' discomfort with ratings. Finally, this results of this study suggest that relationships among class size, grade distributions, and ratings needs further exploration to understand how class size affects the nature and quality of instruction in order to understand better what ratings express and how users are to apply that knowlege to decision-making.

References:

Aleamoni, L.M. (1981). Student ratings of instruction. In J. Millman (ed.) *Handbook of Teacher Evaluation*, Beverly Hills, CA: Sage.

Aleamoni, L.M. (1987). Student rating myths versus research facts. *Journal of personnel evaluation in education, 1* (1), 111-119.

Centra, J. A. (1979). *Determining Faculty Effectiveness.* San Francisco: Jossey Bass, p.44-45 .

Cohen, H. (1984), Inflated grades, deflated courses: has insecurity induced compromise? *Change, 16* (4) 8-10.

Cohen, P.A. (1981). Student ratings of instruction and student achievement: a meta-analysis of multi-section validity studies. *Review of Educational Research, 31* (3), 281-309.

Costin, F. (1978). Do student ratings of college teachers predict student achievement? *Teaching of Psychology, 5 (*2), 86-88.

Costin, F. Greenough, W.T. and Menges, R.J. (1971). Student ratings of college teaching: reliability, validity, and usefulness. *Review of Educational Research, 41* (5), 511-536.

Doyle, K.O., Jr. (1975). Student Evaluation of Instruction. Lexington, MA: D.C. Heath and Company.

Feldman, K.A. (1976). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education, 4*, 69-111.

Franklin, J. and Theall, M. (1989). Who reads ratings: knowledge, attitude and practice of users of student ratings of instruction . Paper presented at the Annual Meeting of the American Eduational Research Association (San Francisco).

Goldman, L. (1985) The betrayal of the gatekeepers: grade inflation. *The Journal of General Education, 37* (2), 97-121.

Machina, K. (1987). Evaluating student evaluations. *Academe, 73* (3) 19-22.

Marsh, H.W. and others (1976). The relationship between students' evaluation of instruction and expected grades. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisico, 1976 (ERIC ED 126140.

Marsh, H.W. (1987). Students evaluations of university teaching: research findings, methodological issues, and directions for future research. *International Journal of Education Research, 11*, 253-388

Martinson, D.L., and Ryan, M. (1984). Student evaluations of faculty should be published cautiously. *Journalism Educator*, Spring, 11-14.

McKeachie, W.J. (1979). Student ratings of faculty: a reprise. *Academe*, October, 384-397.

Milton, O., Pollio, H.R, Eison, J.A. (1986). *Making sense of college grades.* San Francisco, CA: Jossey Bass.

Myers, C. (1979). Student ratings of professors exert growing influence. *The Chronicle of Higher Education*, October 29, p.12.

Nelson, J. P. and Lynch, K.A., Grade inflation, real income, simultaneity, and teaching evaluations. *Journal of Economic Education*, Winter 1984, 21-37.

Seldin , P. (1987). "Q & A" evaluating teaching performance: answers to common questions. AAHE Bulletin, 40 (1), 9-12.